Breast Cancer Classification Through Transfer Learning with Vision Transformer, PCA, and Machine Learning Models

Juan Gutierrez-Cardenas
Carrera de Ingeniería de Sistemas, Universidad de Lima, Lima-Perú

***Abstract*—Breast cancer is a leading cause of death among women worldwide, making early detection crucial for saving lives and preventing the spread of the disease. Deep Learning and Machine Learning techniques, coupled with the availability of diverse breast cancer datasets, have proven to be effective in assisting healthcare practitioners worldwide. Recent advancements in image classification models, such as Vision Transformers and pretrained models, offer promising avenues for breast cancer imaging classification research. In this study, we employ a pretrained Vision Transformer (ViT) model, specifically trained on the ImageNet dataset, as a feature extractor. We combine this with Principal Component Analysis (PCA) for dimensionality reduction and evaluate two classifiers, namely a Multilayer Perceptron (MLP) and a Support Vector Machine (SVM), for breast mammogram image classification. The results demonstrate that the transfer learning approach using ViT, PCA, and an MLP classifier achieves an average accuracy, precision, recall, and F1-score of 98% for the DSMM dataset and 95% for the INbreast dataset, considering the same metrics which are comparable to the current state-of-the-art.**

***Keywords*—Breast cancer; vision transformer; transfer learning; PCA; machine learning**

I. Introduction

Breast cancer, as defined by the World Health Organization (WHO) [39], encompasses a spectrum of diseases. It can manifest as a slow progression without symptoms, or it can take on an aggressive form, invading surrounding tissues and potentially spreading to nearby lymph nodes or other organs. Early identification of breast cancer is of utmost importance to prevent adverse outcomes. As per the National Cancer Institute (NIH) [26], breast cancer ranks as the second most common cause of mortality in the United States. The screening process is essential for the early detection of breast cancer cases prior to the manifestation of symptoms, with mammography being the predominant screening method. Aside from mammography, there are several other techniques available for detecting breast cancer, such as breast ultrasound, breast magnetic resonance imaging (MRI), and biopsy [5].

Classical mammographic images, as seen in the DDSM dataset (Heath et al., 1998), have inherent limitations in terms of image contrast and quality when compared to alternative techniques such as Magnetic Resonance Imaging (MRI). The problem is more noticeable in samples obtained from young women, as their breast tissue density is higher [4].

In order to overcome these constraints, alternative mammographic techniques, such as full-field digital mammography (FFDM), have been developed and are employed to extract information to be used in datasets such as INbreast [27]. The benefits of employing this technique encompass aspects such as patient satisfaction, simplicity of image manipulation, enhanced display contrast, superior detection efficiency, and minimal vulnerability to noise. One significant benefit is that these images can be employed for computer-aided diagnosis (CAD) tools [25].

Among these techniques, the availability of public datasets, particularly those derived from diagnostic mammograms or breast MRI, has facilitated the application of diverse Machine Learning and Deep Learning models for the identification and classification of breast cancer across different stages of the disease. In Tsochatzidis et al. [38], for instance, the researchers used a modified CNN with U-Net-derived image segmentation and evaluated it using the DDSM dataset. In terms of the AUC metric, the authors' diagnostic performance was 0.898. Min et al. [24] employed a Mask R-CNN for mass detection and segmentation using the INbreast dataset in a different study. The average true positive rate that the researchers were able to obtain in this study was 0.9. Readers interested in the utilization of these datasets through the application of Convolutional Neural Networks (CNNs) are encouraged to examine the research conducted by Zhu et al. [40].

In their study Samee et al. [32] used the INbreast and mini-MAIS datasets to demonstrate the efficacy of a breast cancer detection system. The system employed image pre-processing techniques, specifically contrast-limited adaptive histogram equalization (CLAHE) and pixel-wise intensity adjustment, to generate pseudo-colored images. Transfer learning was utilized in conjunction with various deep learning models, including AlexNet, VGG, and GoogleNet, to leverage pre-trained features. Additionally, Logistic Regression and Principal Component Analysis (PCA) were employed to extract the most informative features. The authors applied PCA to mitigate multicollinearity issues that could arise from synthetic image generation. The proposed approach resulted in 23 principal components. Multiple machine learning methods, such as Support Vector Machines (SVM), decision trees, and Convolutional Neural Networks (CNN), were utilized as classifiers. Notably, the CNN classifier achieved the best performance, attaining an accuracy of 98.8% for the MIAS dataset and 98.62% for the INbreast dataset.

Al-Tam et al. [1] used various deep learning models that were employed for both a two-class classifier (benign and malignant) and a three-class classifier that included a normal state as an additional class. The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) and the Digital Database of Screening Mammography (DDSM) datasets were utilized for evaluation. The authors utilized pre-trained models such as VGG16, ResNet50, and ImageNet. Furthermore, they compared the performance of these pre-trained models with a CNN trained from scratch and a hybrid model combining ResNet50 with a Vision Transformer (ViT). Notably, the proposed approach achieved exceptional results with 100% F1-Score, accuracy, and AUC for the binary classification task. However, it is important to consider that these results might be influenced by the quality of information available in the CBIS-DDSM dataset. In the multiclass scenario, the performance metrics decreased to 96% on the validation set and 95% on the test set. The authors acknowledged the need for further evaluation using additional datasets such as INbreast and MAIS to assess the generalizability of their proposed approach.

In their study, Houssein et al. [14] introduced an enhanced version of the Marine Predators algorithm (MPA) called the Improved Marine Predators algorithm (IMPA). This algorithm, which incorporates Opposition-based Learning (OBL), was utilized for hyperparameter optimization of various CNN models on the DDSM and MIAS datasets. Specifically, the authors applied IMPA to optimize the hyperparameters of a ResNet50 model, which employed transfer learning and data augmentation techniques. Notably, the proposed approach achieved compelling results on both datasets. For the CBIS-DDSM dataset, the ResNet50 model attained an accuracy of 98% and an F1-score of 97%. Similarly, on the MAIS dataset, the model achieved an accuracy of 98% and an F1-score of 97%. However, the authors recognized certain limitations of their approach, e.g., the computational cost associated with IMPA was relatively high. Additionally, the proposed architecture was specifically tailored to the tested datasets, which may limit its generalizability.

In Table I, we have summarized the mentioned studies along with others, considering their methodology.

Our main contribution lies in the design of a transfer learning-based Vision Transformer (ViT) that incorporates PCA for feature reduction, addressing the challenge posed by the large number of features extracted from images. This approach is combined with a simple machine learning technique to aid in the classification of breast cancer image samples. The ViT serves as a feature or characteristic extractor from images in our design, and we reduce their dimensionality using PCA to overcome processing time complexity. PCA is commonly used as a pre-processing technique to enhance the efficiency of Machine Learning models [21] by removing unnecessary or irrelevant data [29]. Furthermore, it has demonstrated favourable results in the categorization of breast mammograms [28]. Following this, a simple and non-computationally expensive machine learning technique is employed, with the hypothesis that it will produce accurate results considering the DSSM and INBreast breast cancer image datasets that are comparable to the existing literature. In summary, we aim to leverage the feature extraction capabilities of a state-of-the-art model, such as ViT, and subsequently reduce the dimensionality of these features using PCA. Considering the advantages mentioned before of this dimensionality reduction technique, we then plan to employ computationally non-costly machine learning models like MLP and SVM. Moreover, our current work contributes a proof-of-concept showing how cutting-edge models, like ViT, can be combined with traditional techniques like PCA and machine learning models to produce reliable classification results for breast cancer diagnosis that are on par with those documented in the literature.

TABLE I. RELATED WORKS AND THEIR METHODOLOGY

Authors	Methodology
Tsochatzidis et al. [38]	Employs a modified CNN architecture that incorporates a U-Net for image segmentation during input.
Min et al. [24]	Grayscale images are converted into pseudo-color and the masses are amplified for utilization in a Mask R-CNN that utilizes transfer learning.
Samee et al. [32]	Images are improved through the application of contrast-limited adaptive histogram equalization (CLAHE). A CNN model, selected from AlexNet, VGG, or GoogleNet, is used to extract features, while a Logistic Regression model with PCA is employed for classification.
Al-Tam et al. [1]	The authors employed VGG16, ResNet50, and Imagenet for both binary and multiclass classification. For the final test, they utilized a ResNet50 model that was trained from scratch, in addition to a ViT model.
Samee et al. [33]	The authors utilized pre-trained convolutional neural network (CNN) models, specifically AlexNet, GoogleNet, and VGG-16. The authors utilized pre-trained convolutional neural network (CNN) models, specifically AlexNet, GoogleNet, and VGG-16. The researchers used several feature selection methods, such as Pearson Correlation Coefficient, Cosine Coefficient (mostly used for texts), Euclidean Distance (though Liu and Zhang (2016) warned that it might not be the best way to represent data characteristics, which could lead to poor learning), and Mutual Information. The chosen characteristics were subjected to classification using an ensemble of learners utilizing Discriminant Analysis, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naive Bayes. Nevertheless, it is still uncertain whether they employed a combination of machine learning models or determined which one produced the most optimal outcomes.
Jabeen et al. [19]	The authors utilized a haze-reduced local-global image enhancement technique. The images were subjected to augmentation, and a pre-trained EfficientNet-b0 model was used as a feature extractor, excluding the last three layers. The process of selecting features was conducted utilizing the Equilibrium-Jaya controlled Regula Falsi algorithm. An ensemble of K-nearest neighbors (EKNNs) was utilized for classification.
Our Proposal	The ViT model is utilized as a feature extractor, PCA is employed for dimensionality reduction, and MLP and SVM are used as classifiers for the purpose of comparison.

We have organized our work into the following sections: Section II provides an overview of Transfer Learning and Vision Transformers, covering the relevant theoretical aspects. In Section III, we outline our methodology, describing the algorithms used to guide our procedures; we also had the experimental setup conducted on two breast cancer datasets. Section IV presents the key findings and results obtained from our models. In Section V, we engage in a comprehensive discussion of the results, drawing comparisons with relevant studies that have also explored breast cancer classification. Finally, we conclude our article with a summary of the main insights and conclusions derived from this study in Section VI.

II. BACKGROUND

A. Breast Cancer Datasets: DDSM and INbreast

Multiple Breast Cancer Datasets are available, with some being freely accessible and others requiring permissions for use. This study will employ the DDSM and INbreast datasets, which will be briefly described.

1) DDSM dataset: This dataset is a well-known collection of digitized copies derived from images taken during a screening exam. Furthermore, it includes carefully selected images curated by professionals, displaying an accurate representation of both benign and malignant instances of breast cancer. An inherent concern with this dataset, despite its widespread utilization throughout the years, is the existence of anomalies in certain images, such as the occurrence of dust or scratches, which necessitate careful consideration [13].

2) The INbreast Dataset was obtained from a university hospital in Portugal and consists of samples from both breast cancer patients and healthy individuals. This dataset offers several benefits, such as including samples obtained from patient screenings, diagnoses made based on abnormalities, and follow-up cases of individuals who underwent some type of treatment. Furthermore, the dataset contains a wide range of observations that can be identified during breast exams, including asymmetries, calcifications, distortions, masses, and nodules. The images were acquired using FFDM equipment, which offers superior image quality in comparison to their DDSM equivalent [27].

B. Transfer Learning

Transfer learning is a technique that enables a model to use the knowledge acquired during the training of a previous model rather than starting the training process from scratch. The fundamental idea is that if a model has learned useful representations or variations on a dataset P1, those representations can be transferred or adapted to improve the learning of a new task P2 [10].

To illustrate this concept, let's consider the ResNet model. This model is often pre-trained on a large dataset such as ImageNet, which contains a vast number of images from various categories. When pre-training ResNet, the model learns to recognize general features and patterns in the images. The later layers of the model, which are responsible for making specific predictions, can be replaced with new layers that are tailored to the target task. The reason for this replacement is that the early layers have already captured general features, while the later layers can be fine-tuned to capture task-specific features for the new dataset [20].

For example, if we have a deep learning model based on VGG16, we can exclude the classifier part by disabling or removing the top layer. By doing so, we obtain a feature vector of 4096 numbers. This vector can be serialized and stored, serving as input to a new model. Alternatively, we can replace the classifier part with a new set of convolutional layers if we want to use a different classifier. This adaptability allows us to customize the model architecture according to the specific requirements of the task at hand [3, 8].

C. Vision Transformer

A Vision Transformer (ViT) is a type of attention model initially developed for Natural Language Processing (NLP) tasks but has also shown promise in image analysis. Unlike traditional convolutional neural networks (CNNs), ViT requires fewer computational resources when pre-trained on a large image dataset and subsequently applied to smaller datasets for classification tasks.

The ViT model operates by dividing an image into a set number of patches, each with a fixed size. These patches are then embedded, creating a sequence of embeddings that is subsequently fed into a Transformer Encoder. The Transformer Encoder is made up of self-attention heads and MLP (Multi-Layer Perceptron) blocks, which help the model find patterns and connections in the image [7]. A schematic representation of the ViT model is presented in Fig. 1.

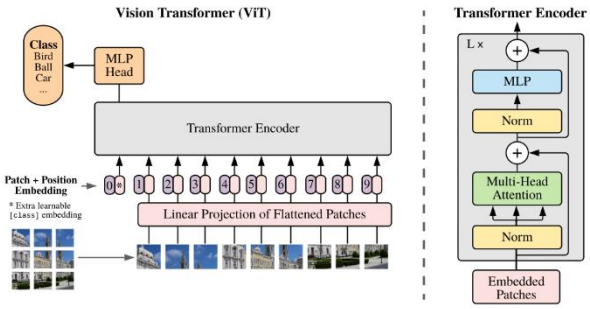


Fig. 1. Vision transformer model [7].

D. Evaluation Metrics

For validating the result or our proposed model we have used the following metrics [9]:

Accuracy: This value represents the proportion of correctly classified instances. It is calculated by considering instances that are predicted to have positive or negative values and belong to one of those classes. The formula is as follows:

$$acc = \frac{1}{|Te|}\sum_{x \in Te} I[\hat{c}(x) = c(x)] \quad (1)$$

In Eq. (1), the Te refers to the test set, while the function $I[x]$ refers to the indicator function. This function takes a value of 1 when the value is correctly classified and 0 in other cases.

Precision: This metric refers to the calculation of the proportion of accurate positive predictions. This means that if the model predicts a value in the positive class, it must be in that class. The formula is as follows:

$$Prec = \frac{TP}{TP+FP} \quad (2)$$

where, TP and FP represent True Positive and False Positive, respectively. A True Positive is an instance that has this value and was correctly classified as positive, whereas a False Positive is an instance that was incorrectly classified as positive but has a negative value.

Recall: This metric refers to the percentage of all positive instances that are correctly predicted. This means that if all positive instances of a model are considered, this metric tells us how many the model correctly predicted. The formula for this metric is as follows:

$$\text{Rec} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

False Negative (FN) refers to instances that are classified as negative but belong to a positive class.

F1-Score: When we want to calculate the average of the incorrect classifications made while considering the set of classes, we can use the F1-score formula:

$$F1 - \text{Score} = \frac{2}{\frac{1}{\text{Prec}}+\frac{1}{\text{Rec}}} \quad (4)$$

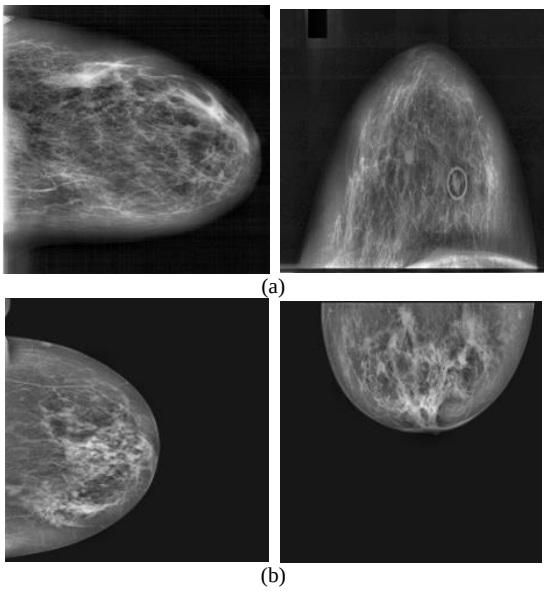
III. Materials and Methods

A. Dataset

The objective of this study was to evaluate the performance of a couple of machine learning models, specifically the Multi-Layer Perceptron (MLP) and Support Vector Classifier (SVC), in conjunction with transfer learning techniques and a Vision Transformer for breast cancer image classification. We conducted experiments using a dataset comprising images of benign and malignant breast samples obtained from the Digital Database for Screening Mammography (DDSM) and the INbreast datasets.

In this study, we collected a dataset of breast mammography images from the Dataset of Breast Mammography Images with Masses (Huang and Lin, 2020), which is available at https://data.mendeley.com/datasets/ywsbh3ndr8/2.
Specifically, we utilized the Digital Database for Screening Mammography (DDSM) and the INbreast datasets [12] from this repository.

The dataset used in this study was compiled from multiple sources. Initially, Huang and Lin [15] selected 106 images from the INbreast dataset, 53 images from the MIAS dataset, and 2188 images from the DDSM dataset. To address the issue of overfitting, a data augmentation technique was employed, which involved multi-angle rotation, flipping, and 11-angle rotation in both horizontal and vertical directions. The compiled dataset, available at https://data.mendeley.com /datasets/ywsbh3ndr8/2, is organized into four folders: DDSM dataset, INbreast dataset, INbreast+MIAS+DDSM dataset, and MIAS dataset. For our experiments, we focused on the DDSM dataset and the INbreast dataset. The DDSM dataset consists of both benign and malignant masses, with 5970 and 7158 samples, respectively. The INbreast dataset contains 2520 samples of benign cases and 5112 samples of malignant cases. An example of both types of samples from these datasets is shown in Fig. 2.



(a) (b)

Fig. 2. A couple of samples from the benign and malignant masses as found in the DDSM (a) and INbreast datasets (b).

Considering the number of samples, we can observe that the DDSM dataset consists of 55% malignant masses and 45% benign masses, while the INbreast dataset consists of 33% malignant cases and 67% benign cases. Based on these percentages, we can conclude that the data is not imbalanced. It is worth noting that the study conducted by Haibo and García [11] suggested that a dataset can be considered imbalanced if the minority class constitutes less than 10% of the total samples. It is important to mention that their study focused on dichotomous classes, similar to the ones examined in our research. In scenarios in which the data is unbalanced, techniques such as data augmentation [15] can be used. This includes image rotation at 11 angles in both horizontal and vertical directions, ranging from 30º to 330º degrees in 30-degree increments. Additionally, horizontal and vertical flipping can be used.

B. Methodology

In the Fig. 3, we have depicted the steps followed in our work and that can be summarize in the following steps:

Step 1: We obtained a collection of images from the DDSM and INbreast datasets that represent both benign and malignant formations related to breast cancer. Prior to being inputted into a Vision Transformer (ViT) model, this data undergoes resizing and normalization.

Step 2: The ViT model operates as a feature extractor through the utilization of transfer learning. To carry out the

mentioned function, the classifier head is detached from this model.

Step 3: After that, a PCA model receives the features that the ViT model generated. During this stage, we isolate a subset of components that possess the ability to elucidate the majority of the data. The objective is to decrease the dimensionality of the data, rendering it more manageable for straightforward and computationally efficient machine learning models such as a Multilayer Perceptron (MLP) and a Support Vector Machine (SVM).

Step 4: The hyperparameters of both models are adjusted, and the classification results are assessed using metrics such as accuracy, precision, recall, and F1-score.

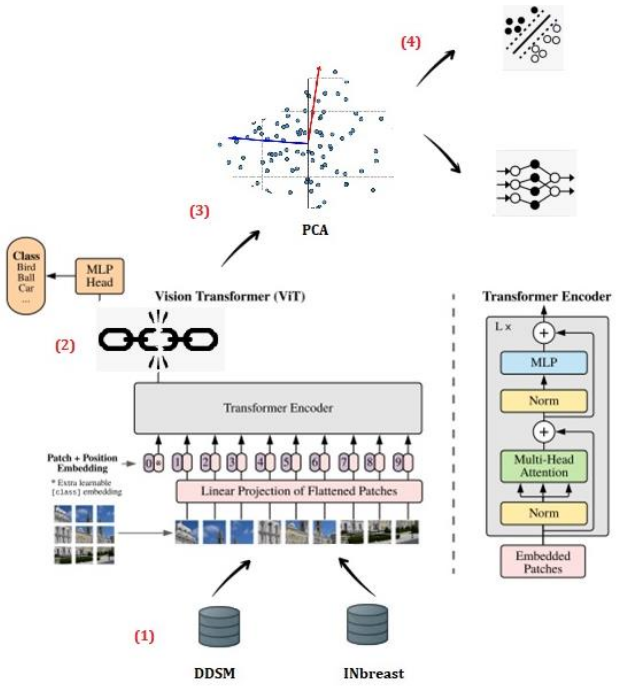


Fig. 3. The integration of a ViT model as a feature extractor, coupled with PCA and machine learning models as classifiers, serves to identify benign and malignant cases of breast cancer (figure of the ViT obtained from the work of [7].

C. Experimentation

Considering the DDSM Dataset, which contains both benign and malignant images of the breast, we performed image resizing to 224 x 224 pixels and channel normalization with a value of 0.5 [5]. Subsequently, we read the images from their respective folders and assigned a label of 0 for benign masses and 1 for malignant masses. To determine the number of epochs for model optimization, we employed a standard split of 80% for training and 20% for validation, with a batch size of 32. We utilized the pre-trained Vision Transformer (ViT) model "facebookresearch/deit:main" with the "deit_base_patch16_224" architecture, which was pre-trained on the ImageNet-1k dataset at a resolution of 224 x 224 with fixed patches of 16 x 16 [36, 16]. For the loss function, we chose Cross Entropy, a commonly used metric for estimating probabilities in breast cancer classification [17, 18]. The optimizer selected was Adam, supported by previous studies [18, 34], with a learning rate of 0.001. It is important to note that we employed this configuration to create a feature extraction model by removing the last layer, which served as a classifier, after the training phase. Additionally, we flattened our data into a 2D tensor, where each row corresponds to the features extracted from an image. With this setup, we conducted experiments and obtained the training and validation loss curves, as shown in Fig. 4. To evaluate the model's performance, we tested various numbers of epochs and decided to maintain a value of five based on the results obtained.

For the INbreast dataset, the methodology was similar, with the exception that when we plotted the loss curves for a fixed number of epochs using the same learning rate as applied to the DDSM dataset, we observed that the validation set's loss did not decrease significantly. This indicated the occurrence of overfitting. To address this issue, we manually adjusted the learning rate and determined that a value of 0.0001 resulted in a rapid decrease of the validation set's loss with eight epochs.

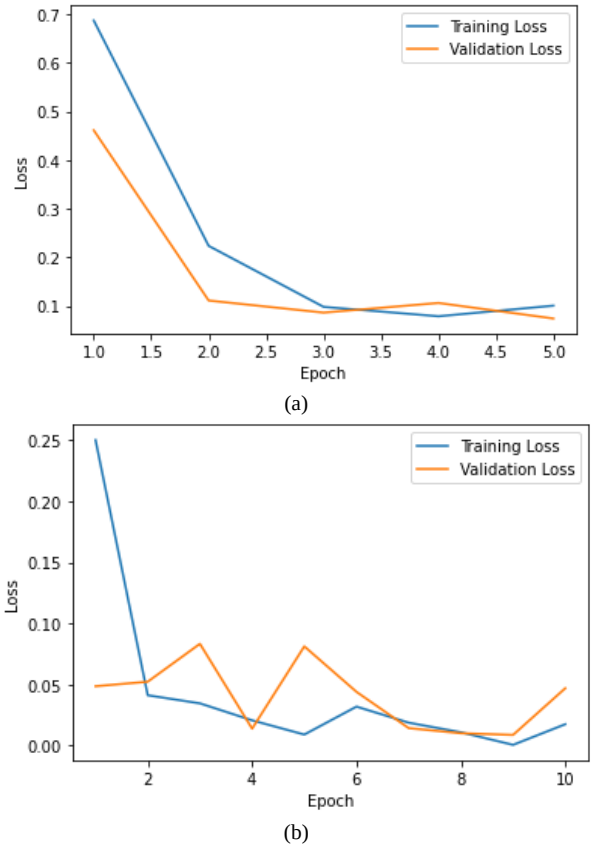


Fig. 4. Training and validation curve losses for obtaining the number of epochs for the DDSM (a) and INbreast datasets (b).

Algorithm 1 outlines the steps performed in the study:

1) Define the dataset directory and the image transformation pipeline using PyTorch's `transforms.Compose()` function.

2) Create a custom dataset class that is inherited from PyTorch's Dataset class. In the constructor, initialize the root directory, transformation pipeline, and targets. Implement the `__len__` method to return the total number of samples in the

dataset. Implement the `__getitem__` method to compute the number of samples in each class, determine the class of the current sample based on its index, load the corresponding image and label, apply the transformation pipeline to the image, and return the transformed image and label.

3) Split the dataset into training and validation sets using PyTorch's `random_split()` function.

4) Define data loaders for the training and validation sets using PyTorch's DataLoader class.

5) Load the pre-trained ViT model from Facebook Research using the `torch.hub.load()` function.

6) Define the loss function as the cross-entropy loss and choose the optimizer as Adam.

7) Train the model on the training set for a specified number of epochs. During training, fine-tune the pre-trained ViT model on the custom dataset, enabling it to learn task-specific features. Use the loss function and optimizer to update the model's parameters.

8) After each epoch, validate the model on the validation set and calculate the validation loss using the loss function. This step monitors the model's performance on unseen data and helps prevent overfitting.

9) Save the trained visual transformer model to a file using the `torch.save()` function.

10) Plot the training and validation losses using Matplotlib. This visualization aids in tracking the model's performance during training and identifying any issues, such as overfitting.

When the model was trained, given the number of epochs obtained, we decided to use PCA for dimensionality reduction. The choice of PCA was mainly because we wanted to reduce the number of features given a certain number of components. In the experiments performed, we found that the number of components that explained 95% of the data was 43 for the DDSM dataset, while the number of suitable components found for the INbreast dataset was of 1933 components. It is not surprising that INbreast required a greater number of principal components. We hypothesize that the main reason for this is that the dataset consists of electric signals converted into images, which provides more detailed information than the DDSM dataset [34].

An algorithm is provided for utilizing the saved model from Algorithm 1 to obtain the desired number of components using PCA. The components will serve as input features in the machine learning model:

Algorithm 2:

1) Load the trained visual transformer model by invoking the function load_visual_transformer().

2) Define a feature extractor by removing the classification head from the pre-trained ViT model through the creation of a torch.nn.Sequential() object.

3) Utilize a data loader to apply the feature extractor to the images in the dataset. For each batch of images, extract the features using the feature extractor, flatten the resulting feature maps, and store the features and corresponding labels in separate lists.

4) Concatenate the feature vectors and labels, transforming them into numpy arrays.

5) Apply PCA to the feature vectors to reduce their dimensionality.

At this stage, we opted to employ two machine learning models: a multilayer perceptron (MLP) and a support vector classifier (SVC). For the MLP model, we explored the following number of hidden layers as a hyperparameter grid:

Hidden layer size 1: A single hidden layer with the same number of neurons as the input features.

Hidden layers size 64: In this case, we employed two hidden layers. The first layer had the same number of neurons as the input, and the second layer had 64 neurons.

Hidden layers size 128: Like the configuration mentioned earlier, but with the second hidden layer having 128 neurons.

Hidden layers size 256: Again, similar to the previous configurations, with the number of neurons in the hidden layer now set to 256.

This grid served as input for a grid search cross-validation function that utilized five folds to determine the best number of hidden layers as a hyperparameter for this model. After applying the grid search function, we identified the best hyperparameter values to be 256 neurons for the DDSM dataset and 128 neurons for the INbreast dataset, respectively.

For the DDSM dataset, we utilized two hidden layers with 256 and 128 neurons, respectively. Meanwhile, for the INbreast dataset, we employed the same number of hidden layers but with 1933 and 128 neurons in each layer. The activation function used was ReLU.

Concerning the hyperparameters for the SVC, we utilized a hyperparameter grid consisting of the following values:

C (penalization factor): 0.01, 0.1, 1, 10, 100

Kernels: Linear, Polynomial, RBF

Gamma value (only applicable to RBF): 0.001, 0.01, 0.1

Subsequently, we performed a grid search cross-validation with five folds to obtain the best hyperparameters, considering the grid.

For the DDSM dataset, the SVC model was configured with an RBF kernel, a penalty parameter C of 100, and a gamma value of 0.001. As for the INbreast dataset, the hyperparameters for the SVC model were set as follows: RBF kernel, C of 100, and gamma value of 0.1.

IV. Results

After employing transfer learning using a Vision Transformer (ViT) as described in the methodology section and applying the aforementioned classifier methods, we obtained the following results for both models that are presented in Table II.

TABLE II. METRICS OBTAINED FROM THE MODELS EVALUATED IN THE DDSM (A) AND INBREAST (B) DATASETS

(a)

ViT model + Classifier using PCA (DDSM)/Metric	Acc	Prec	Recall	F1-score
MLP	0.9819	0.983	0.9836	0.983
SVC	0.9672	0.962	0.9786	0.970

(b)

ViT model + Classifier using PCA (INbreast)/Metric	Acc	Prec	Recall	F1-score
MLP	0.943	0.954	0.9624	0.9582
SVM	0.843	0.810	1	0.8953

Table II displays the metrics for Accuracy, Precision, Recall, and F1-score of our proposed model, which were calculated using the DDSM and INbreast datasets. It is noteworthy that our Precision and Recall results consistently exceed 97% on average. The precision and recall in medical diagnosis are of utmost importance. In this cancer situation, the recall metric prioritizes false negatives, while precision focuses on false positives. Furthermore, the f1-score, which is the harmonic mean that takes into account both precision and recall, has yielded an average of 96%. Upon evaluating both models, it can be deduced that they exhibit minimal occurrences of false positive and false negative predictions. Furthermore, they provide a well-balanced performance in terms of precision and recall.

From the data presented in Table II, it is evident that the MLP model outperformed the SVC classifier for both datasets. It is worth noting that a five-fold cross-validation was utilized for validating our results in each model.

It is worth noting that the MLP model outperformed the SVM model on both datasets. In the case of the DDSM dataset, the difference between the two models is only about one point. However, the distinction observed in the INbreast dataset is more pronounced, with a difference of nearly 10 points in Accuracy and Precision between both models.

According to the authors [22], both datasets, DDSM and INbreast, were subjected to data augmentation techniques such as rotation and flipping. The difference between the original data from both datasets (i.e., data that had not been augmented) was significant. For the DDSM dataset, 2188 images were augmented to 13128, while 106 mass images were augmented to 7632 for the INbreast dataset. The proportion of augmented data in the INbreast dataset far outnumbers that in the DDSM dataset.

At this point, we can speculate that this disparity may have contributed to the SVM model's lower results compared to its MLP counterpart, which demonstrated a greater ability to generalize its classification capabilities in both datasets. These findings are intriguing, especially considering the work of Shen et al. [34]. Their research indicated that the INbreast dataset, containing FFDM (full-field digital mammography) images with varied intensity profiles, allowed them to evaluate the suitability of a particular classifier model across several mammography platforms. This property stems from the fact that FFDM images replace X-rays with electrical signals, allowing them to be reproduced across multiple devices [35].

V. DISCUSSION

In comparison to other studies that have utilized the DDSM dataset, we did not find any previous work that employed transfer learning combined with PCA and machine learning models. In the study conducted by Ayana et al. [2], transfer learning was also utilized with various ViT models, such as Swim and Pyramid, along with image augmentation to address the issue of dataset imbalance. Their results, obtained when training the models from scratch, ranged from a 78% in accuracy, precision, and F1-score. Furthermore, the authors explored CNN models including ResNet, EfficientNet, and InceptionNet, achieving an average accuracy, F1-score, and recall of 94%. We believe it is important to mention recall as a crucial metric since the consequences of missing or misclassifying a cancer screening can be detrimental. However, our approach, incorporating PCA, transfer learning, and machine learning models, yielded promising results with an average performance of 98% across the evaluated metrics.

Another study conducted by Ragab et al. [30] investigated two datasets, namely the DDSM and the Curated Breast Image Subset of the DDSM (CBIS-DDSM). The authors employed image enhancement techniques, including Contrast-Limited Adaptive Histogram Equalization (CLAHE), to improve image definition. They also performed image segmentation and utilized data augmentation. It is worth mentioning that CLAHE was also applied to the DDSM dataset with data augmentation, as employed in our present study. Ragab et al. [30] performed feature extraction using a Deep Convolutional Neural Network (DCNN), specifically AlexNet, which was pre-trained on the ImageNet dataset. Their combined model, consisting of the DCNN and a Support Vector Machine (SVM) classifier, achieved an accuracy of 87.2% using a medium Gaussian kernel function. Although there are some differences between the datasets used in their study and ours, there are notable similarities, such as both datasets being based on the DDSM dataset and the utilization of similar techniques for image preprocessing and augmentation. The authors' use of image segmentation is a distinct difference from our approach.

In the research conducted by Salama et al. [31], two datasets, namely the DDSM and the curated DDSM, were utilized. The authors applied data augmentation techniques such as rotation and employed two deep learning models, ResNet-50 and VGG-16. Transfer learning was performed from the ImageNet dataset, and the classification layer was modified to accommodate only two classes. Although it appears that both models were used as feature selection algorithms, no explicit mention of this approach was found. The outputs from both models were then used as inputs for an SVM classifier. While the authors mentioned hyperparameter tuning for the deep learning methods employed, there was no information provided regarding hyperparameter tuning for the SVM model. The results obtained for the DDSM dataset using the VGG-16 model yielded an average accuracy, AUC, sensitivity, precision, and F1-score of 94%. For the CBIS-DDSM dataset, the VGG-16 and ResNet-50 models combined with the SVM classifier and five-fold cross-validation achieved an average accuracy, AUC, sensitivity, precision, and F1-score of 96% for the former and 95% for the latter, in addition to an average F1-score of 93%.

Other researchers, such as Tsochatzidis [37], conducted experiments with various deep learning models, including AlexNet, VGG-16, VGG-19, ResNet-50, ResNet-152, GoogLeNet, and Inception-BN. In their study, the authors initialized the weights of their models from scratch and also applied transfer learning techniques to the DDSM-400 and CBIS-DDSM datasets. Data augmentation was not employed in their experiments. According to their findings, training the models from scratch yielded the best performance with AlexNet, achieving an accuracy of 62% for the DDSM-400 dataset and 65% for the CBIS-DDSM dataset. However, the best results were obtained when using pre-trained initialized weights for both datasets. In particular, the ResNet-based model achieved an accuracy of 85% for the DDSM-400 dataset and an average accuracy of 80% for the CBIS-DDSM dataset.

Das et al. [6] conducted a study where they evaluated the performance of various deep learning models on breast cancer datasets, including CBIS-DDSM and INbreast. Their experiments involved both shallow neural networks and deep neural networks. Among the models tested, the Xception network demonstrated the best performance, achieving an accuracy of 89% for CBIS-DDSM and 95% for INbreast. The authors suggested that the higher accuracy obtained on the INbreast dataset could be attributed to the higher image quality compared to CBIS-DDSM.

As of the writing of this article, we have not found any existing research that has utilized a transfer-learning model based on Vision Transformer (ViT) in conjunction with Principal Component Analysis (PCA) for dimensionality reduction. Furthermore, our results show that a simple Multilayer Perceptron (MLP) model with two hidden layers, employed as a classifier, outperforms SVM-based approaches. We strongly believe that leveraging pre-trained models, particularly those based on attention mechanisms like ViT, in combination with dimensionality reduction techniques applied to the data, holds promise for achieving superior performance. Moreover, these approaches can be beneficial in scenarios where computational resources for data processing are limited.

Table III, which is an expansion of Table I mentioned in the Introduction section, contains the datasets utilized in the reviewed studies, as well as the metrics derived from the outcomes of the various applied models.

TABLE III. COMPARISON OF OTHER STUDIES WITH OUR CURRENT PROPOSAL

Authors	Methodology	Dataset used	Results
Tsochatzidis et al. [38]	Employs a modified CNN architecture that incorporates a U-Net for image segmentation during input.	DDSM-400 and CBIS-DDSM	AUC 89.8% and 86.2%
Min et al. [24]	Grayscale images are converted into pseudo-color and the masses are amplified for utilization in a Mask R-CNN that utilizes transfer learning.	Inbreast	90% (True Positive Rate) TPR
Samee et al. [32]	Images are improved through the application of contrast-limited adaptive histogram equalization (CLAHE). A CNN model, selected from AlexNet, VGG, or GoogleNet, is used to extract features, while a Logistic Regression model with PCA is employed for classification.	Inbreast and MIAS	98.8% of accuracy using MIAS and 98.6% using MIAS.
Al-Tam et al. [1]	The authors employed VGG16, ResNet50, and Imagenet for both binary and multiclass classification. For the final test, they utilized a ResNet50 model that was trained from scratch, in addition to a ViT model.	CBIS-DDSM	100% in the F1-score for the binary classification, other metrics in average made a 96% in the multiclass classification.
Samee et al. [33]	The authors utilized pre-trained convolutional neural network (CNN) models, specifically AlexNet, GoogleNet, and VGG-16. The researchers applied a series of feature selection techniques, including Pearson Correlation Coefficient, Cosine Coefficient (mainly used for texts), Euclidean Distance (although Liu and Zhang [23] warned about possible drawbacks in representing data characteristics, which could result in suboptimal learning), and Mutual Information. The chosen characteristics were subjected to classification using an ensemble of learners utilizing Discriminant Analysis, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naive Bayes. Nevertheless, it is still uncertain whether they employed a combination of machine learning models or determined which one produced the most optimal outcomes.	Inbreast	98.06% in Sensitivity and 98.5% in Accuracy.
Jabeen et al. [19]	The authors utilized a haze-reduced local-global image enhancement technique. The images were subjected to augmentation, and a pre-trained EfficientNet-b0 model was used as a feature extractor, excluding the last three layers. The process of selecting features was conducted utilizing the Equilibrium-Jaya controlled Regula Falsi algorithm. An ensemble of K-nearest neighbors (EKNNs) was utilized for classification.	CBIS-DDSM Inbreast	Average Accuracy of 95.4% and 99.7%
Our Proposal	The ViT model is utilized as a feature extractor, PCA is employed for dimensionality reduction, and MLP and SVM are used as classifiers for the purpose of comparison.	DDSM Inbreast	Average Accuracy, Precision, Recall and F1-score of 98% for the DDSM dataset with MLP as classifier. The same metrics give us an average of 95.4% for the InBreast dataset.

Regarding the limitations identified in the current research, it is important to note that the author faced difficulties in finding appropriate examples to aid in the coding process of the Vision Transformer (ViT) when utilized as a feature extractor instead of a classifier. Although there is limited literature on using ViT models as feature extractors, we are still confident in their potential for applications where the features obtained can be used as input for Machine learning (ML) models. ML models provide several benefits, such as their interpretability, decreased computational requirements, and ongoing potential usefulness in the domain of medical diagnosis.

An important upcoming task would be to evaluate various Vision Transformer (ViT) models in combination with different subsets of machine learning (ML) models, such as Random Forest or other boosting-based methods, to ascertain if these model combinations can enhance the reported results. In addition, performing experiments with diverse datasets, e.g., the MIAS or the BancoWeb Lapimo, datasets beyond those specified in the present study, would allow for the assessment of the overall efficacy of an integrated model that includes ViT, dimensionality reduction of features, and machine learning techniques in diagnosing breast cancer scenarios using mammography data.

VI. Conclusion

The classification of samples obtained from mammograms holds utmost importance, as early detection of malignant masses can significantly impact patient outcomes. In this study, we demonstrated the efficacy of a transfer learning model based on Vision Transformer (ViT), coupled with Principal Component Analysis (PCA) for feature reduction, and a simple Multilayer Perceptron (MLP) model. Our results were found to be comparable to existing literature that employs Convolutional Neural Network (CNN) models based on transfer learning in conjunction with deep learning models. These findings highlight the potential of using ViT-based transfer learning approaches, combined with dimensionality reduction techniques and simple Machine Learning classifiers, to achieve accurate mammogram classification results.

References

[1] Al-Tam RM, Al-Hejri AM, Narangale SM, Samee NA, Mahmoud NF, Al-masni MA, et al. 2022. A Hybrid Workflow of Residual Convolutional Transformer Encoder for Breast Cancer Classification Using Digital X-ray Mammograms. Biomedicines.10:2971. doi:10.3390/biomedicines10112971.

[2] Ayana, G., K. Dese, Y. Dereje, Y. Kebede, H. Barki, D. Amdissa, N. Husen, F. Mulugeta, B. Habtamu, and S.-W. Choe. 2023. Vision-Transformer-Based Transfer Learning for Mammogram Classification. Diagnostics 13, no. 2 (January 4): 178. doi:10.3390/diagnostics13020178.

[3] Brownlee J. 2019. Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python. 198-200. Machine Learning Mastery.

[4] Bhushan, A., A. Gonsalves, and J.U. Menon. 2021. Current State of Breast Cancer Diagnosis, Treatment, and Theranostics. Pharmaceutics 13, no. 5 (May 14): 723.

[5] Centers for Disease Control and Prevention (CDC). Breast Cancer. Accessed 11, November 2023. https://www.cdc.gov/cancer/breast/basic_info/diagnosis.htm.

[6] Das, H.S., A. Das, A. Neog, S. Mallik, K. Bora, and Z. Zhao. 2023. Breast Cancer Detection: Shallow Convolutional Neural Network against Deep Convolutional Neural Networks Based Approach. Frontiers in Genetics 13 (January 4): 5:37. doi:10.3390/jimaging5030037.

[7] Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, et al. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. Accessed November 11, 2023. arXiv. http://arxiv.org/abs/2010.11929.

[8] Ferguson, M., R. Ak, Y.-T.T. Lee, and K.H. Law. 2017. Automatic Localization of Casting Defects with Convolutional Neural Networks. In 2017 IEEE International Conference on Big Data (Big Data), 1726–1735. Boston, MA: IEEE. doi:10.1109/BigData.2017.8258115.

[9] Flach, P. 2012. Machine learning: The art and science of algorithms that make sense of data, 300-54.USA: Cambridge University Press. https://doi.org/10.1017/CBO9780511973000.

[10] Goodfellow I., Bengio Y., Courville A. 2016. Deep Learning. 539-540. MIT Press.

[11] Haibo, H., and Garcia, E.A. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21, no. 9 (September): 1263–1284. doi:10.1109/TKDE.2008.239.

[12] Heath, M., K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran. 1998. Current Status of the Digital Database for Screening Mammography. In Digital Mammography, ed. N. Karssemeijer, M. Thijssen, J. Hendriks, and L. Van Erning, 13:457–460. Computational Imaging and Vision. Dordrecht: Springer Netherlands. http://link.springer.com/10.1007/978-94-011-5318-8_75.

[13] Heath, M., Bowyer, K., Kopans, D., Moore, R., and Kegelmeyer, W. P. 2001. The Digital Database for Screening Mammography. In Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., 212-218. Medical Physics Publishing. ISBN 1-930524-00-5.

[14] Houssein EH, Emam MM, Ali AA. 2022. An optimized deep learning architecture for breast cancer diagnosis based on improved marine predators algorithm. Neural Comput & Applic.34:18015–33. doi:10.1007/s00521-022-07445-5.

[15] Huang, M.-L., and T.-Y. Lin. Dataset of Breast Mammography Images with Masses. 2020. Data in Brief 31: 105928. doi: 10.17632/ywsbh3ndr8.2.

[16] Hugging Face. "facebook/deit-base-patch16-224." Hugging Face, n.d. Accessed November 11, 2023. https://huggingface.co/facebook/deit-base-patch16-224.

[17] Jaamour, A. 2020. Breast Cancer Detection in Mammograms Using Deep Learning Techniques. MSc. diss., University of St. Andrews.

[18] Jaamour, A., C. Myles, A. Patel, S.-J. Chen, L. McMillan, and D. Harris-Birtill. 2023. A Divide and Conquer Approach to Maximise Deep Learning Mammography Classification Accuracies. PLOS ONE 18, no. 5 (May 26): e0280841. doi:10.1371/journal.pone.0280841.

[19] Jabeen, K., M.A. Khan, J. Balili, M. Alhaisoni, N.A. Almujally, H. Alrashidi, U. Tariq, and J.-H. Cha. 2023. BC2NetRF: Breast Cancer Classification from Mammogram Images Using Enhanced Deep Learning Features and Equilibrium-Jaya Controlled Regula Falsi-Based Features Selection. Diagnostics 13, no. 7 (March 25): 1238.

[20] Keheller J. 2019. Deep Learning. 236-237. MIT Press.

[21] Kherif, F., and A. Latypova. 2020. Principal Component Analysis. Machine Learning, 209–225. Elsevier. https://linkinghub.elsevier.com/retrieve/pii/B9780128157398000122.

[22] Lin, T. and M. Huang. (2020), Dataset of Breast mammography images with Masses. Accesed December 24, 2023. https://data.mendeley.com/datasets/ywsbh3ndr8/5.

[23] Liu, M., and D. Zhang. 2016. Feature Selection with Effective Distance. Neurocomputing 215 (November): 100–109.

[24] Min, H., D. Wilson, Y. Huang, S. Liu, S. Crozier, A.P. Bradley, and S.S. Chandra. 2020. Fully Automatic Computer-Aided Mass Detection and Segmentation via Pseudo-Color Mammograms and Mask R-CNN. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 1111–1115. Iowa City, IA, USA: IEEE. https://ieeexplore.ieee.org/document/9098732/.

[25] Muller, S. 1999. Full-Field Digital Mammography Designed as a Complete System. European Journal of Radiology 31, no. 1 (July): 25–34.

[26] National Cancer Institute (NIH). "Breast Cancer Screening (PDQ®)–Patient Version." National Cancer Institute. Accessed December 28, 2023. https://www.cancer.gov/types/breast/patient/breast-screening-pdq

[27] Moreira, I.C., I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, and J.S. Cardoso. 2012. INbreast. Academic Radiology 19, no. 2 (February): 236–248.

[28] Oral, C., and H. Sezgin. 2013. Effects of Dimension Reduction in Mammograms Classification. In 8th International Conference on Electrical and Electronics Engineering (ELECO), 630–633. Bursa, Turkey: IEEE. http://ieeexplore.ieee.org/document/6713912/.

[29] Palo, H.K., S. Sahoo, and A.K. Subudhi. 2021. Dimensionality Reduction Techniques: Principles, Benefits, and Limitations. Data Analytics in Bioinformatics, ed. R. Satpathy, T. Choudhury, S. Satpathy, S.N. Mohanty, and X. Zhang, 77–107. 1st ed. Wiley. https://onlinelibrary.wiley.com/doi/10.1002/9781119785620.ch4.

[30] Ragab, D.A., M. Sharkas, S. Marshall, and J. Ren. 2019. Breast Cancer Detection Using Deep Convolutional Neural Networks and Support Vector Machines. PeerJ 7 (January 28): e6201. doi:10.7717/peerj.6201. 11.

[31] Salama, W.M., A.M. Elbagoury, and M.H. Aly. 2020. Novel Breast Cancer Classification Framework Based on Deep Learning. IET Image Processing 14, no. 13 (November): 3254–3259. doi:10.1049/iet-ipr.2020.0122.

[32] Samee NA, Alhussan AA, Ghoneim VF, Atteia G, Alkanhel R, Al-antari MA, et al. 2022a. A Hybrid Deep Transfer Learning of CNN-Based LR-PCA for Breast Lesion Diagnosis via Medical Breast Mammograms. Sensors;22:4938. doi:10.3390/s22134938.

[33] Samee, N.A., G. Atteia, S. Meshoul, M.A. Al-antari, and Y.M. Kadah. 2022b. Deep Learning Cascaded Feature Selection Framework for Breast Cancer Classification: Hybrid CNN with Univariate-Based Approach. Mathematics 10, no. 19 (October 4): 3631.

[34] Shen, L., L.R. Margolies, J.H. Rothstein, E. Fluder, R. McBride, and W. Sieh. 2019. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Scientific Reports 9, no. 1 (August 29). doi:10.1038/s41598-019-48995-4.

[35] Stanford Medicine. https://stanfordhealthcare.org/medical-tests/m/mammogram/digital-mammography.html. Accessed, December 24, 2023.

[36] Touvron, H., M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training Data-Efficient Image Transformers & Distillation through Attention. Last Modified 2021. Accessed November 11, 2023. arXiv. http://arxiv.org/abs/2012.12877.

[37] Tsochatzidis, L., L. Costaridou, and I. Pratikakis. 2019. Deep Learning for Breast Cancer Diagnosis from Mammograms—A Comparative Study. Journal of Imaging 5, no. 3 (March 13): 37. doi:10.3390/jimaging5030037.

[38] Tsochatzidis, L., P. Koutla, L. Costaridou, and I. Pratikakis. 2021. Integrating Segmentation Information into CNN for Breast Cancer Diagnosis of Mammographic Masses. Computer Methods and Programs in Biomedicine 200 (March): 105913.

[39] World Health Organization (WHO). Breast Cancer. Accessed November 11, 2023. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[40] Zhu, Z., S.-H. Wang, and Y.-D. Zhang. 2023. A Survey of Convolutional Neural Network in Breast Cancer. Computer Modeling in Engineering & Sciences 136, no. 3: 2127–2172..